Further Comment

The Rasch Model Revisited*

Thomas O'Shea
simon fraser university
David F. Robitaille
university of british columbia

In our article on the development of a test-item bank in mathematics utilizing the one-parameter simple logistic model (Robitaille and O'Shea, 1983), we described the various developmental and analytic procedures which were followed in the Mathematics Achievement Test (MAT) Project in British Columbia. This included a relatively brief, non-technical description of our reasons for utilizing what is commonly referred to as the Rasch model, and an overview of some of the issues surrounding the use of this procedure. Our article was not designed to provide a thorough review of those issues, and inclusion of such an analysis in an article of that kind would have been inappropriate. We did point out that those who planned to use this technique should familiarize themselves with the arguments for and against its use, and references were made to the relevant literature.

We had anticipated that our article would generate some reaction and we would have welcomed thoughtful and constructive criticism of the different aspects of the project. This turned out to be a vain hope. In their reaction paper, McLean and Ragsdale (1983) elected to focus solely on our use of the Rasch model and, as seems frequently to be the case in discussions of that model, they contributed a great deal of heat but very little light. Their strident tone and intemperate language were unusual to see in a journal like *CJE/RCE*. Unfortunately however, they failed to address the substantive issues regarding the methodology employed in the MAT Project.

We, along with many others, do not believe that employment of the Rasch model is "inappropriate in the past, inappropriate today, inappropriate tomorrow" (McLean and Ragsdale, 1983). The choice of a model in

* This comment is made on a previously published article by Leslie D. McLean and Ronald G. Ragsdale, "The Rasch model for achievement tests – Inappropriate in the past, inappropriate today, inappropriate tomorrow," *Canadian Journal of Education/Revue canadienne de l'education*, 1983, *8*(1), 71–76.

educational measurement is dependent upon the circumstances in which it is to be employed and its demonstrable utility.

In their paper, McLean and Ragsdale are highly critical of our use of the expression "within reasonable limits" in discussing the properties of the Rasch model. Such qualifiers are commonly employed in the statement of propositions and generalizations in mathematics, including statistics. We used the phrase to temper some of the claims made in the past by overly zealous proponents of the model. For example, samples of persons should be drawn from the population which has been exposed to instruction on the material being examined; such samples are "within reasonable limits." One sample has to be chosen to provide a centre for the scale, and that may require probability sampling. Other than that, the samples do not have to be random, as they must be for the development and norming of traditional tests.

McLean and Ragsdale suggest that the classical test-theory notion of equivalent forms may be used to arrive at the same type of "item-free" ability estimates obtained using the Rasch model. We suggest that this is not the case since parallel forms require equal means and equal variances, whereas the latent trait procedure allows the use of sets of items with differing mean difficulty levels. Moreover, the construction of classically equivalent forms is difficult, expensive, and impractical for general achievement tests.

McLean and Ragsdale express concern about the issue of unidimensionality and the use of a single scale across strands at each grade level, that is, "fitting all types of items to the same scale." One may hypothesize any number of possible underlying abilities, but the empirical demonstration of those abilities is another matter. We found that very few items, between 2 and 3%, were deleted because they did not fit the model. More were deleted, between 8 and 9%, because they did not meet the criteria traditionally employed to select multiple-choice items. Furthermore, when the remaining items were re-calibrated, a common structure was *found* to exist; it was not created. Our position is that any item in the bank may be used as an indicator of a student's level of mathematical achievement after that student has been exposed to the usual curriculum. For example, the rank order of students on the geometry items will be similar to that on the consumer mathematics strand, for those students who have been taught both sets of content, because a common mathematical ability underlies the scale.

Each item in the pool was calibrated on a representative sample of students selected from across the province, a sample which may therefore be considered as representative of the entire range of levels of instruction, of competency of teachers, of ability of students, and of variations in curriculum. When tests are produced by selecting items from the pool for a particular content area, those tests are intended to measure the achievement of students on that content, not on any other. If a general

indication of grade-level achievement were desired, a random sample of items might be selected from the bank to produce a test for that purpose. If, however, a particular teacher has not dealt with, say, geometry to the same degree as the majority of other teachers in the province, or if certain students have not understood the material presented, it would be expected that such students would not perform as well as others on a test designed to measure achievement in geometry. It is not "within reasonable limits" to suggest that the model purports to predict achievement at the classroom level on one set of items by using an unrelated set of items. A distinction has to be kept in mind between the calibration phase and the testing phase.

McLean and Ragsdale also criticize our use of the Rasch model because they believe that meeting the assumptions of the model is critical. This is an issue which must be addressed empirically by investigating the robustness of the model with respect to violations of those assumptions. We contend that the end product justifies the procedures used in our case. In their critique, McLean and Ragsdale failed to deal with that end product – i.e. the validation phase of the MAT Project.

Perhaps the significance of those results went unnoticed. To recapitulate, the calibration data on the items from each Grade 7/8 Strand Test were used to predict a raw score boundary value for each of the five letter grades *A*, *B*, *C*, *P*, and *F*. The percentage of students falling into each letter-grade classification was determined during the calibration phase by using teachers' estimates of their students' achievement in mathematics. A separate norming sample of 1000 students for each Strand Test was selected to determine the legitimacy of the predicted raw score boundaries. Because the mean difficulties of items on the various Strand Tests differed, the boundary values for letter-grade levels varied across tests. The validation sample confirmed the predicted cutoff levels within the error of measurement associated with each score. An examination of Table 1[1] in our article (Robitaille and O'Shea, 1983) shows that the raw score boundaries differed considerably across strands.

In our article, we suggested that the results of the validation process indicated several potential benefits. One of these has been realized recently as the B.C. Ministry of Education has mandated the use of summative tests in certain subjects at different grade levels as of June 1983. The existence of the item bank in mathematics will make it possible to produce reliable and valid tests, interpretable by classroom teachers, without having to spend a great deal of time and money on the traditional norming phase of test development.

We concur with McLean and Ragsdale that the use of the Rasch model for multiple-choice questions is open to question. At the time of the development of our item bank in British Columbia (1978–79), it was not feasible to employ any of the other latent trait models. In recent years there has been considerable progress in the estimation procedures for

more complex latent trait models, particularly for the three-parameter model which includes guessing and discrimination parameters, as well as one for item difficulty. The application of latent trait theory has become an accepted part of testing as shown by the use of the Rasch model in such diverse applications as the Portland Public Schools project, the Los Angeles County Test Development Centre, the Georgia Regents' Testing Program, the KeyMath Diagnostic Arithmetic Test, and the Woodcock Reading Mastery Tests. The three-parameter model is being used in the California State Assessment, by CTB/McGraw-Hill, and by the Educational Testing Service.

As an alternative to the use of the Rasch model, McLean and Ragsdale put forward a proposal based on their work with the Ontario Assessment Instrument Pool (OAIP). As they admit, they did not have sufficient space in which to adequately describe their project and one wonders why they bothered to mention it at all. If the OAIP alternative is a fruitful one, we suggest they publish an adequate description of it so that interested parties may evaluate its validity and effectiveness.

NOTE

[1] In our article, there is an apparent discrepancy between the grade-level boundaries shown in Table 1 and those shown in Figure 5. The values in Figure 5 are actually those from the Sets and Numbers Strand Test for Grades 3/4, not for Grades 7/8.

REFERENCES

McLean, L. D., & Ragsdale, R. G. The Rasch model for achievement tests – Inappropriate in the past, inappropriate today, inappropriate tomorrow. *Canadian Journal of Education*, 1983, *8*(1), 71–76.

Robitaille, D. F., & O'Shea, T. The development of an item bank in mathematics using the Rasch model. *Canadian Journal of Education*, 1983, *8*(1), 57–70.